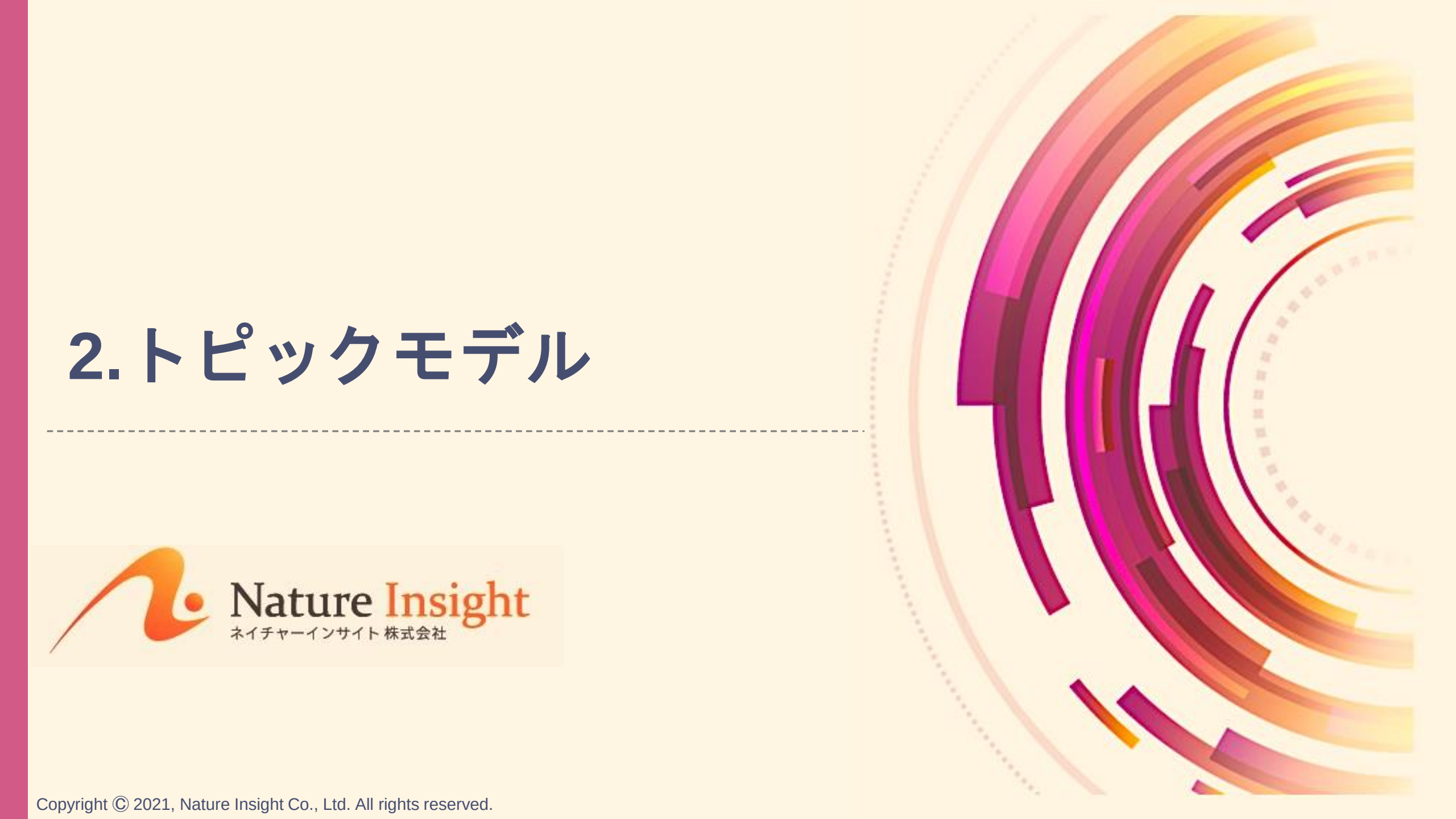
2. トピックモデル



Nature Insight
ネイチャーインサイト 株式会社

目次

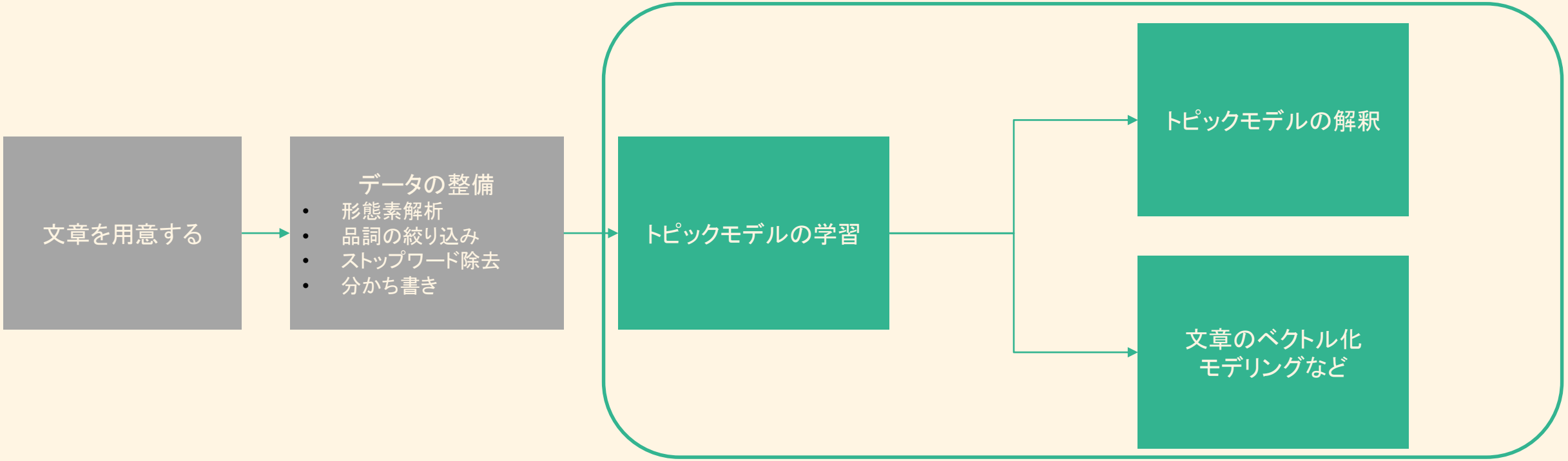
この講義の目的

講義の目的

今回の講義では、文章の次元削減・クラスタリング手法である「トピックモデル」を取り扱います。

データの整備は取り扱いませんが、この部分は既知として進めますので、

もし不安な方は自然言語処理の基礎講義の受講で復習してからこの講義を受けてください。



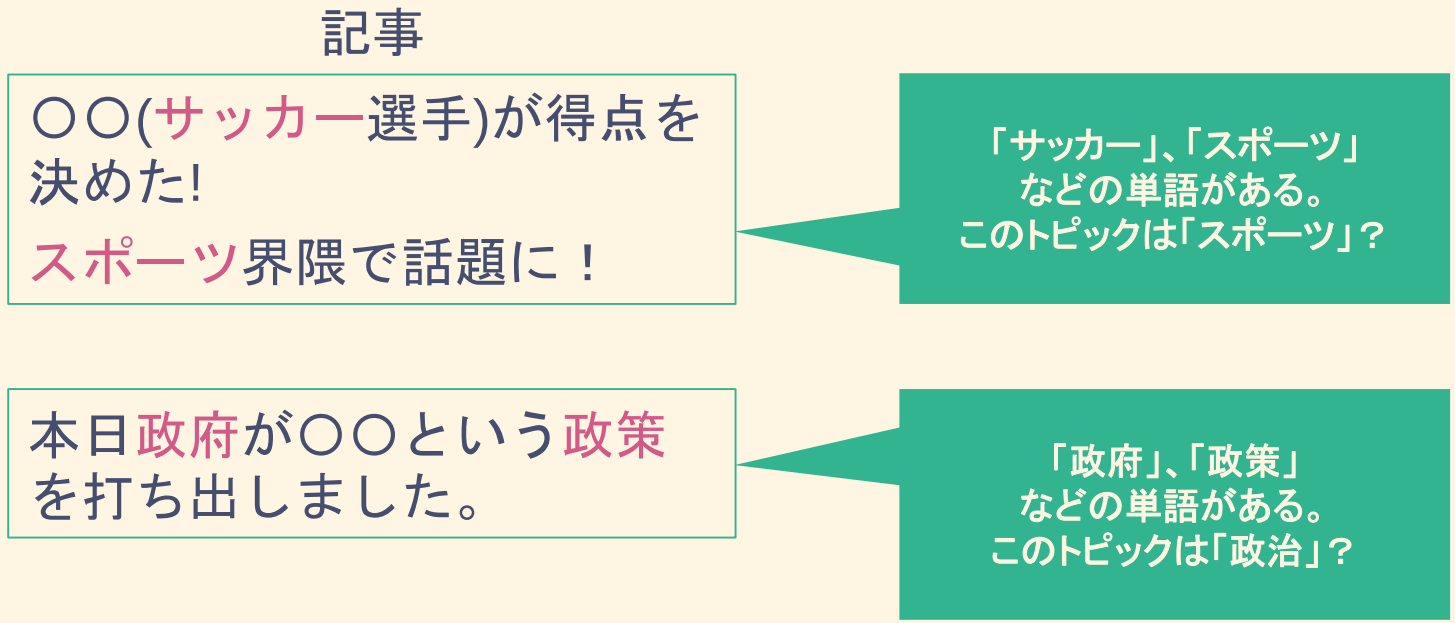
1.トピックモデルの概要と種類

トピックモデルの概要と種類

トピックモデルの概要

例えばあなたがニュースの記事を分類するとします。
その際、記事の単語から「スポーツ」、「政治」、「音楽」などのトピック(意味のまとまり)で分類できることに気が付くと思います。

このように、文章にはトピックがあるとして、このトピックをもとに文章を分類する手法を「トピックモデル」といいます。



トピックモデルの概要と種類

Nature Insight

トピックモデルの種類

トピックモデルには大きく分けて3種類あります。

- LSA(潜在的意味解析)
  … 特異値分解(SVD)を利用して次元削減を行い、文章中の情報量を凝縮する
- PLSA
  …LSAを確率モデルとして解釈したもの。トピックや単語を確率分布として表現できる。
- LDA(潜在的ディリクレ配分法)
  …PLSAと考え方は同じだが、PLSAをベイズ化し、新たなデータに対して頑健にしたもの。

今回の講義では、gensimで実装されているLSAとLDAについて解説していきます。

2. LSA(潜在的意味解析)

LSA(潜在的意味解析)

LSAの概要1

LSA(潜在的意味解析)は、自然言語処理・情報検索分野*で開発された手法で、**「美味しい」と「美味い」などの似た意味を表す単語を同じとみなして分類・検索を行う**ことを目的とします。

以下はそれぞれの文章のBag-of-Wordsです。

	寿司	美味しい	美味い	車	買う
doc1：寿司美味しい	1	1	0	0	0
doc2：寿司美味い	1	0	1	0	0
doc3：車を買う	0	0	0	1	1

「美味しい」、「美味い」は別単語ですが、doc1、doc2はほとんど同じことを言っています。
これを似ているとみなすにはどうすればよいでしょうか？

* 情報検索分野ではLSIと呼ばれます。

LSA(潜在的意味解析)

Nature Insight

LSAの概要2

LSAでは、文章のベクトルを単語の頻度で表す代わりに、**トピック**で表します。
(以降、「**文章ごとのトピックの重み**」と呼びます。)

	Topic1	Topic2
doc1：寿司美味しい	1.225	0
doc2：寿司美味い	1.225	0
doc3：車を買う	0	1.414

LSAでは、**類義性、多義性を考慮して**分類されます。(doc1、doc2はともにTopic1)。
また、**次元数が単語からトピックになり**、次元削減に成功しています。

さらに、LSAでは**トピックごとの単語の重み**も出力されます。

	寿司	美味しい	美味い	車	買う
Topic1	0.816	0.408	0.408	0	0
Topic2	0	0	0	0.707	0.707

Topic1では「寿司」「美味しい」「美味い」が、Topic2では「車」「買う」がTopicの話題であることが予想されます。

LSA(潜在的意味解析)

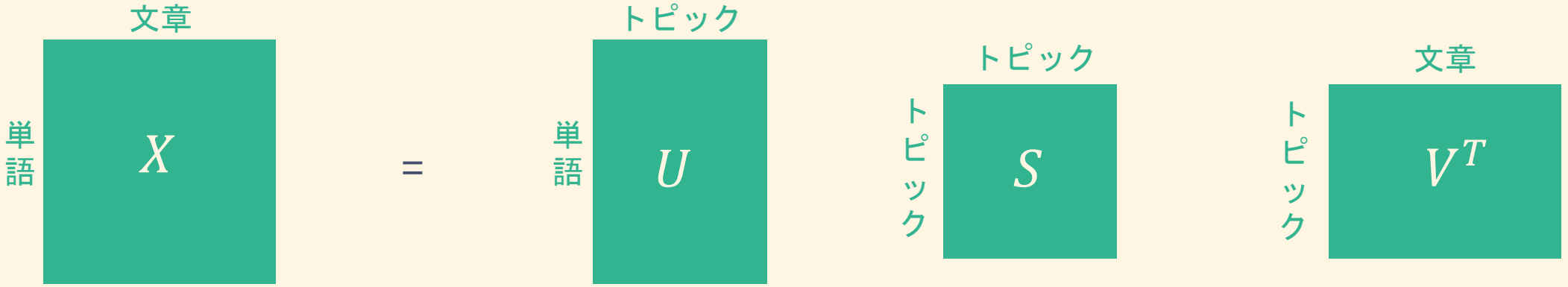
LSAの仕組み1

LSAでは、**特異値分解(SVD)**を使用して次元削減を行っています。

特異値分解*とは、行列Xを列直交行列U, Vと対角成分を持つ行列Sを用いて、以下のように分解することです。

ここでは、mは単語数、 $k(= rank(X))$はトピック数、 nは文章数に該当します。

$$X = U S V^T$$



Bag-of-words

$m \times n$　　$m \times k$　　$k \times k$　　$k \times n$

* U: $m \times m, S: m \times n, V: n \times n$とする場合もあります。

LSA(潜在的意味解析)

LSAの仕組み2

LSAでは、Sの特異値の絶対値の大きい順に指定のトピック数k'個まで減らし、次元削減を行います。(トピック数は分析者が決定します。)

$$\tilde{X} = \tilde{U}\,\tilde{S}\,\tilde{V}^T$$

次元削減後の行列

$m \times n$

トピックごとの
単語の重み

$m \times k'$

$k' \times k'$

$k' \times n$

実際にトピックごとの単語の重みは$\tilde{U}$、次元削減後の文章ごとのトピックの重みを作成するには、$\tilde{U}^T X$を計算します*。

次のページで、次元削減の具体例を解説します。 * $\tilde{S}\tilde{V}^T$を文章ごとのトピックの重みにする場合もあります。

LSA(潜在的意味解析)

LSAの具体例1

実際に特異値分解+次元削減(トピック=2)を行うと以下のようになります。

X^T ：Bag-of-Words

	寿司	美味しい	美味い	車	買う
doc1：寿司美味しい	1	1	0	0	0
doc2：寿司美味い	1	0	1	0	0
doc3：車を買う	0	0	0	1	1

特異値分解+次元削減(トピック=2)

$\tilde{U}^T$ ：トピックごとの単語の重み

	寿司	美味しい	美味い	車	買う
Topic1	0.816	0.408	0.408	0	0
Topic2	0	0	0	0.707	0.707

2×5

$\tilde{S}$

	Topic1	Topic2
Topic1	1.732	0
Topic2	0	1.414

2×2

$\tilde{V}^T$

	Topic1	Topic2
doc1	0.707	0
doc2	0.707	0
doc3	0	1

3×2

LSA(潜在的意味解析)

Nature Insight

LSAの具体例2

$\tilde{U}$をそれぞれのトピックについての単語の重みだと考えれば、$\tilde{U}^T X$は文章ごとのトピックの重みを表すと考えられます。
この変換によって、それぞれの文章の次元が単語数→トピック数になり、次元が削減されます。

$\tilde{U}^T$：トピックごとの単語の重み

	寿司	美味しい	美味い	車	買う
Topic1	0.816	0.408	0.408	0	0
Topic2	0	0	0	0.707	0.707

X^T：Bag-of-Words

	寿司	美味しい	美味い	車	買う
doc1	1	1	0	0	0
doc2	1	0	1	0	0
doc3	0	0	0	1	1

$(\tilde{U}^T X)^T$：文章ごとのトピックの重み

	Topic1	Topic2
doc1	1.225	0
doc2	1.225	0
doc3	0	1.414

文章ベクトルの次元が「単語→トピック」になっている

LSA(潜在的意味解析)

LSAの次元削減結果

Bag-of-Words(X)と、LSAによる次元削減後($\tilde{X}$)の行列を比較してみます。

次元削減前の行列はスパース(0が多い)であるのに対し、次元削減後の行列は共起している単語の0が埋まります。

X^T：Bag-of-Words

	寿司	美味しい	美味い	車	買う
doc1：寿司美味しい	1	1	0	0	0
doc2：寿司美味い	1	0	1	0	0
doc3：車を買う	0	0	0	1	1

LSAで次元削減

「美味しい」・「美味い」の0が埋まっている

$\tilde{X}^T$：LSAによる次元削減後の行列

	寿司	美味しい	美味い	車	買う
doc1：寿司美味しい	1.732	0.866	0.866	0	0
doc2：寿司美味い	1.732	0.866	0.866	0	0
doc3：車を買う	0	0	0	1.414	1.414

LSA(潜在的意味解析)

LSAまとめ

メリット

- 文章ごとのベクトルの成分が単語からトピックになり、次元削減に成功している。
- それぞれのトピックに対してなんらかの意味が付加できるようになっている。
- 「美味しい」、「美味い」のような同じ文脈で使われる言葉は、同じトピックに分けることによって同じような意味として解釈できる。(類義性)
- 「うまい(美味い/上手い)」のような同じ単語でも異なる文脈で使用される言葉は、異なるトピックに分けることによって別の言葉として解釈できる。(多義性)

デメリット

- $\tilde{U}$、$\tilde{V}$ は負の値をとることもあるため、値の解釈が難しい。
- U が列直交行列であるので、トピック同士の単語分布はできるだけ異なるように生成される。(似たようなトピックがあるほうが自然)

LSA(潜在的意味解析)

Nature Insight

LSAの実装

livedoorニュースコーパスを使用してLSAを実装してみましょう。
コードは「2_トピックモデル.ipynb」を参照してください。
livedoorは以下のようなデータで、ラベル(media)がついています。

	url	datetime	title	body	media
0	http://news.livedoor.com/article/detail/5978741/	2011-10-30T10:15:00+0900	【DVDエンター！】誘拐犯に育てられた女が目にした真実は、孤独か幸福か	2005年11月から翌2006年7月まで読売新聞にて連載された、直木賞作家・角田光代による初...	movie-enter
1	http://news.livedoor.com/article/detail/6322901/	2012-02-29T11:45:00+0900	藤原竜也、中学生とともにロケット打ち上げに成功	「アンテナを張りながら生活をしていけばいい」¥n2月28日、映画『おかえり、はやぶさ』（3月...	movie-enter
2	http://news.livedoor.com/article/detail/6176324/	2012-01-09T14:00:00+0900	『戦火の馬』ロイヤル・プレミアにウィリアム王子＆キャサリン妃が出席	3月2日より全国ロードショーとなる、スティーブン・スピルバーグの待望の監督最新作『戦火の馬』...	movie-enter
3	http://news.livedoor.com/article/detail/6573929/	2012-05-19T12:00:00+0900	香里奈、女子高生100人のガチンコ質問に回答「ラーメンも食べる」	女優の香里奈が18日、都内で行われた映画『ガール』（5月26日公開）の女子高生限定試写会にサ...	movie-enter
4	http://news.livedoor.com/article/detail/5914880/	2011-10-05T19:11:00+0900	ユージの前に立ちはだかったJOY「僕はAKBの高橋みなみを守る」	5日、東京・千代田区の内幸町ホールにて、映画『キャプテン・アメリカ/ザ・ファースト・アベンジ...	movie-enter
...	...	...	...	...	...
7362	http://news.livedoor.com/article/detail/6530260/	2012-05-05T09:55:00+0900	好きな戦士を作ってドラゴンボールの世界で天下一武道会優勝だ！「挑戦！天下一武道会」【Andr...	どんな戦士を作るかはユーザー次第！¥n国民的人気を誇る鳥山明氏のマンガ／アニメである「DRA...	smax
7363	http://news.livedoor.com/article/detail/6681611/	2012-06-21T20:55:00+0900	NTTドコモ、GALAXY SIII SC-06DとF-09D ANTEPRIMAの発売日を...	GALAXY SIIIが6月28日、F-09D ANTEPRIMAが6月27日に発売！¥nN...	smax
7364	http://news.livedoor.com/article/detail/6856578/	2012-08-15T11:55:00+0900	NTTドコモ、Android向け「docomo Wi-Fiかんたん接続アプリ」をバージョンア...	shimajiro@mobiler¥nNTTドコモは、同社の公衆無線LANサービス「doco...	smax
7365	http://news.livedoor.com/article/detail/6678539/	2012-06-21T06:55:00+0900	NTTドコモ、PRADA Phone by LG L-02Dのデコメ絵文字popが正常に表示...	PRADA Phone by LG L-02Dにソフトウェア更新！¥nNTTドコモは20日、...	smax
7366	http://news.livedoor.com/article/detail/6869011/	2012-08-20T08:55:00+0900	NTTドコモ、公式オンラインショップでも端末複数台購入で最大10,500円／台の割り引きが受...	NTTドコモは17日、公式オンラインショップ「ドコモオンラインショップ」において端末を複数台...	smax

LSA(潜在的意味解析)

Nature Insight

LSAの結果1

LSAでは、「トピックごとの単語の重み」と、「文章ごとのトピックの重み」が出力されます。
(今回はトピック数=6とした。)
解釈は少ししづらいですが、文章をトピックの特徴量とすることで次元削減に成功しています。

トピックごとの単語の重み

```
[(0,
  '-0.992*"ソフトバンク" + -0.051*"iPhone" + -0.044*"ステージ" + -0.038*"テックランド"'),
 (1,
  '-0.326*"smartphone" + -0.223*"人" + -0.211*"アプリ" + -0.201*"MAX" + -0.196*"画面"'),
 (2,
  '-0.414*"人" + -0.337*"自分" + -0.287*"女性" + 0.264*"ソフトウェア" + -0.225*"映画"'),
 (3,
  '-0.796*"ソフトウェア" + 0.214*"smartphone" + -0.135*"アップデート" + -0.129*"情報"'),
 (4,
  '0.529*"アプリ" + -0.322*"映画" + 0.273*"画面" + 0.190*"iPhone" + -0.167*"ソフトウェア"'),
 (5,
  '-0.548*"映画" + -0.295*"作品" + 0.257*"人" + 0.232*"女性" + 0.209*"自分" + 0.198*"男性"')]
```

1:スマホについての話題？
5:映画についての話題？

文章ごとのトピックの重み

	topic0	topic1	topic2	topic3	topic4	topic5	media
0	-0.047068	-1.892558	-3.392763	-0.308726	-2.173075	-3.377918	peachy
1	-0.078341	-0.581746	-0.627948	-0.010990	-0.769814	-1.260609	peachy
2	-0.035639	-1.177262	-2.241439	-0.358486	0.133030	1.488334	peachy
3	-0.015502	-0.729775	-1.013647	-0.045940	-0.130009	0.538670	peachy
4	-0.005821	-0.303132	-0.367087	0.001161	-0.043065	0.098493	peachy
...	...	...	...	...	...	...	...
7362	-0.032505	-1.016627	-0.000171	0.170272	0.427924	-0.051231	it-life-hack
7363	-0.087204	-2.266165	0.445771	0.411062	0.933109	-0.197487	smax
7364	-0.125168	-3.020753	-0.108339	0.416961	0.792024	-0.306381	it-life-hack
7365	-1.934114	-31.972141	15.985088	11.398469	-12.248043	6.808332	smax
7366	-0.043862	-1.278580	-0.758618	-0.093349	0.534088	-0.891614	kaden-channel

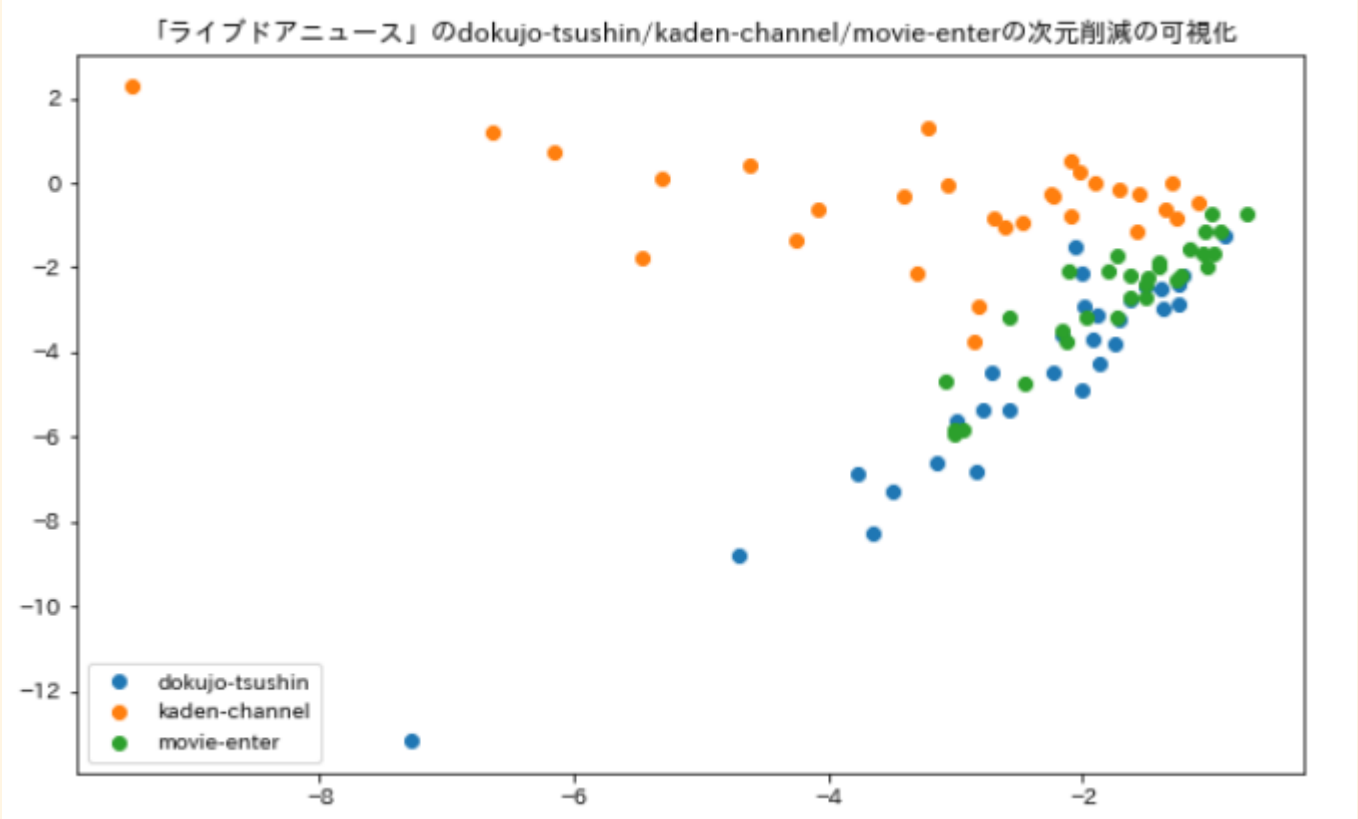
7367 rows × 7 columns

LSA(潜在的意味解析)

LSAの結果2

トピック1をx軸、トピック2をy軸として、dokujo-tsushin/kaden-channel/movie-enterのグラフを書いています。

なんとなくですが、ラベルごとに分かれていることが確認できます。



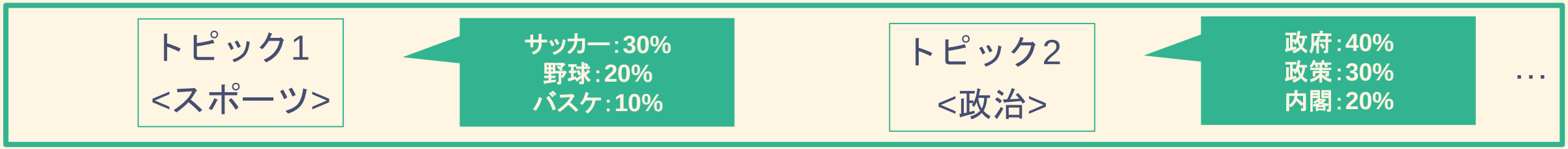
3. LDA(潜在的ディリクレ配分法)

LDA(潜在的ディリクレ配分法)

LDAの概要

もう一度、あなたがニュースの記事を分類するとします。
記事の単語から「スポーツ」、「政治」、「音楽」などのトピックで分類できそうですが、
これを確率分布として表示したいと考えます。(以降「文章-トピック分布」といいます。)

さらに、そのトピックで出現する単語の確率分布も表示します。
(以降を「トピック-単語分布」といいます。)



LDAでは、このように、文章から「文章-トピック分布」と、「トピック-単語分布」を求めて文章を分類します。(トピックの内容は「トピック-単語分布」から分析者が考える必要があります。)

LDA(潜在的ディリクレ配分法)

LDAの仕組み1

LDAでは以下のように、Bag-of-Wordsから「文章-トピック分布」と「トピック-単語分布」の2つの分布を求めることを目的とします。

Bag-of-Words

	寿司	美味しい	美味い	車	買う
doc1：寿司美味しい	1	1	0	0	0
doc2：寿司美味い	1	0	1	0	0
doc3：車を買う	0	0	0	1	1

文章-トピック分布

	Topic1	Topic2
doc1	0.9	0.1
doc2	0.9	0.1
doc3	0.2	0.8

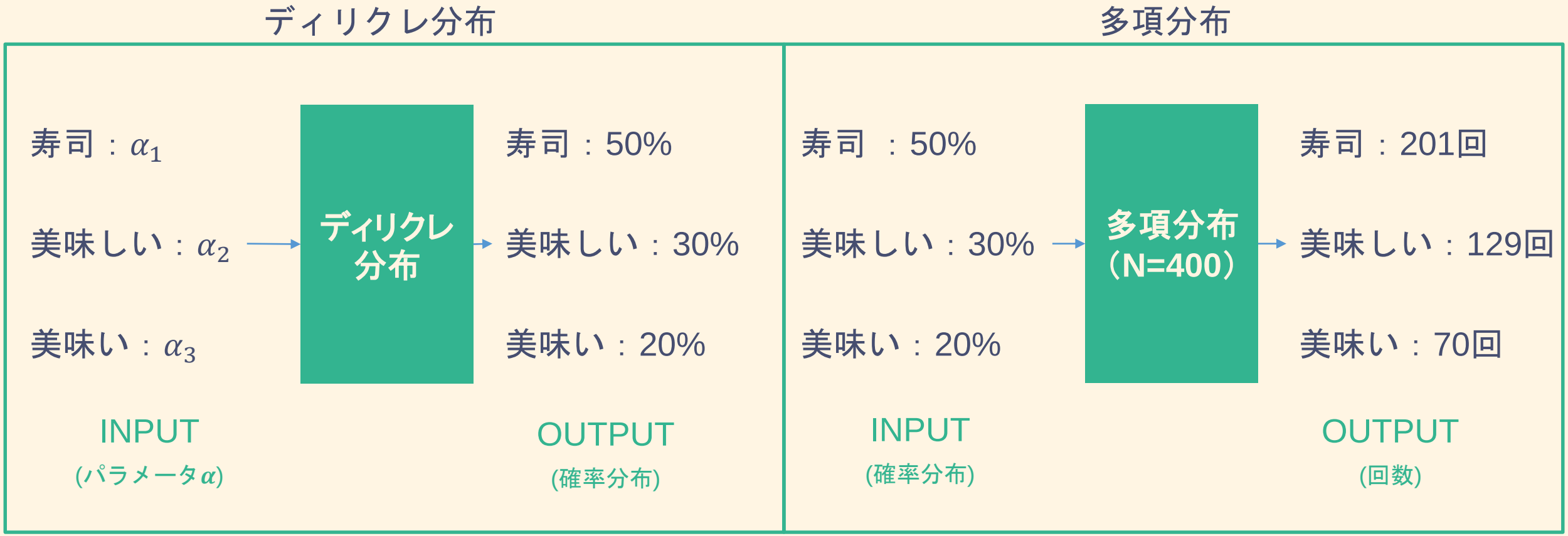
トピック-単語分布

	寿司	美味しい	美味い	車	買う
Topic1	0.4	0.2	0.2	0.1	0.1
Topic2	0.1	0.1	0.1	0.35	0.35

LDA(潜在的ディリクレ配分法)

ディリクレ分布と多項分布

LDAを理解するには、ディリクレ分布と多項分布という確率分布を知ることが必要です。
ディリクレ分布はパラメータαから確率分布を、多項分布は確率分布から回数の分布を出力します。



LDA(潜在的ディリクレ配分法)

LDAのモデル構造1

LDAは以下によってBag-of-Wordsが生成されるモデルを仮定します。

パラメータ

α

文章-トピック分布

	Topic1	Topic2
doc1	0.9	0.1
doc2	0.9	0.1
doc3	0.2	0.8

トピックの回数

	Topic1	Topic2
doc1	181	19
doc2	180	20
doc3	40	160

この分布が
与えられている

Bag-of-Words

	寿司	美味しい	美味い	車	買う
doc1	100	95	5	0	0
doc2	100	5	95	0	0
doc3	0	0	0	100	100

この分布達を知りたい

パラメータ

β

トピック-単語分布

	寿司	美味しい	美味い	車	買う
Topic1	0.4	0.2	0.2	0.1	0.1
Topic2	0.1	0.1	0.1	0.35	0.35

LDA(潜在的ディリクレ配分法)

LDAのモデル構造2

パラメータ

α

ディリクレ分布

文章-トピック分布

	Topic1	Topic2
doc1	0.9	0.1
doc2	0.9	0.1
doc3	0.2	0.8

トピックの回

	Topic1
doc1	181
doc2	180
doc3	

パラメータα, βから
ディリクレ分布を用いて、
確率分布である
文章-トピック分布と、
トピック-単語分布を作成。

					買う
doc1		95	5	0	0
doc2	100	5	95	0	0
doc3	0	0	0	100	100

パラメータ

β

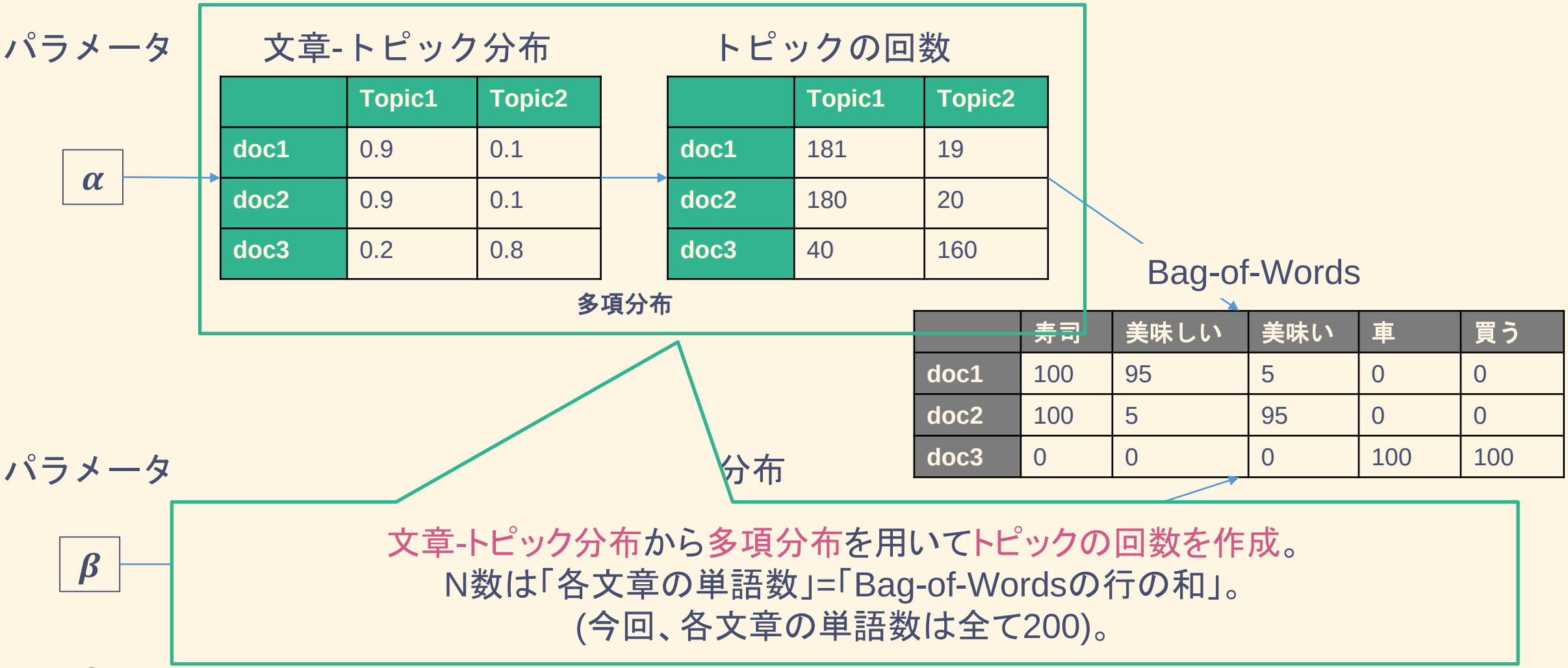
ディリクレ分布

トピック-単語分布

	寿司	美味しい	美味い	車	買う
Topic1	0.4	0.2	0.2	0.1	0.1
Topic2	0.1	0.1	0.1	0.35	0.35

LDA(潜在的ディリクレ配分法)

LDAのモデル構造3



	Topic1	Topic2
doc1	0.9	0.1
doc2	0.9	0.1
doc3	0.2	0.8

	Topic1	Topic2
doc1	181	19
doc2	180	20
doc3	40	160

	寿司	美味しい	美味い	車	買う
doc1	100	95	5	0	0
doc2	100	5	95	0	0
doc3	0	0	0	100	100

LDA(潜在的ディリクレ配分法)

LDAのモデル構造4

トピック-単語分布から多項分布を用いて**Bag-of-Wordsを生成。**

N数はトピックの回数によって分ける。

例) doc1のBag-of-Wordsは、
Topic1：181回、Topic2：19回
「トピック-単語分布」から生成される。

トピックの回数

	Topic1	Topic2
doc1	181	19
doc2	180	20
doc3	40	160

Bag-of-Words

	寿司	美味しい	美味い	車	買う
doc1	100	95	5	0	0
doc2	100	5	95	0	0
doc3	0	0	0	100	100

パラメータ

β

トピック-単語分布

	寿司	美味しい	美味い	車	買う
Topic1	0.4	0.2	0.2	0.1	0.1
Topic2	0.1	0.1	0.1	0.35	0.35

多項分布

LDA(潜在的ディリクレ配分法)

LDAのモデル構造5

Bag-of-Wordsが与えられたとき、最も確率が高くなるようなαとβを求めていきます。
求める方法は、変分ベイズ法やサンプリング近似法などがあります。
もしそれらの詳しい解説が知りたい方は参考文献1を参考にしてください。

LDA(潜在的ディリクレ配分法)

LDAとPLSA(補足)

LDAの前身となるPLSAについても紹介しておきます。

PLSAはLSAの「トピック-単語」と「文章-トピック」の重みを確率分布として表すことを目的としており、LDAの目的とほぼ同義です。

LDAとPLSAの違いは、推定方法です。LDAはベイズ推定法、PLSAは最尤推定法となります。

この性質の違いにより、PLSAでは、LDAに比べて以下の傾向があります。

- 事前分布を仮定しないため、文章のみから「単語-トピック分布」と「文章-トピック分布」を求めることができる。
- 求めるパラメータが多いため、LDAに比べて過学習しやすい。
- 新規文章に対して適用することが困難

自然言語処理では、トピックモデルはLDAで行うことが多いですが、 LDAでうまく分かれない場合や、事前分布を使用したくない場合はPLSAを使用するのも一つの手です。

LDA(潜在的ディリクレ配分法)

LDAまとめ

メリット

- 文章ごとのベクトルの成分が単語からトピックになり、次元削減に成功している。
- 各文章のトピックの確率分布が出る。(LSAでは確率分布になっていなかった。)
  →最も確率の高いトピックを選択することでクラスタリングが可能になる。
- それぞれのトピックに対して単語の確率分布が出る。(LSAでは確率分布になっていなかった。)
  →トピックに対して意味付けを行いやすい。
- (PLSAに比べて)過学習しづらい。

デメリット

- LSAと比較すると学習速度が遅い。

LDA(潜在的ディリクレ配分法)

LDAの実装

livedoorニュースコーパスを使用して実際にトピックモデル(LDA)を実装してみましょう。コードは「2_トピックモデル.ipynb」を参照してください。

	url	datetime	title	body	media
0	http://news.livedoor.com/article/detail/5978741/	2011-10-30T10:15:00+0900	【DVDエンター！】誘拐犯に育てられた女が目にした真実は、孤独か幸福か	2005年11月から翌2006年7月まで読売新聞にて連載された、直木賞作家・角田光代による初...	movie-enter
1	http://news.livedoor.com/article/detail/6322901/	2012-02-29T11:45:00+0900	藤原竜也、中学生とともにロケット打ち上げに成功	「アンテナを張りながら生活をしていけばいい」¥n2月28日、映画『おかえり、はやぶさ』（3月...	movie-enter
2	http://news.livedoor.com/article/detail/6176324/	2012-01-09T14:00:00+0900	『戦火の馬』ロイヤル・プレミアにウィリアム王子＆キャサリン妃が出席	3月2日より全国ロードショーとなる、スティーブン・スピルバーグの待望の監督最新作『戦火の馬』...	movie-enter
3	http://news.livedoor.com/article/detail/6573929/	2012-05-19T12:00:00+0900	香里奈、女子高生100人のガチンコ質問に回答「ラーメンも食べる」	女優の香里奈が18日、都内で行われた映画『ガール』（5月26日公開）の女子高生限定試写会にサ...	movie-enter
4	http://news.livedoor.com/article/detail/5914880/	2011-10-05T19:11:00+0900	ユージの前に立ちはだかったJOY「僕はAKBの高橋みなみを守る」	5日、東京・千代田区の内幸町ホールにて、映画『キャプテン・アメリカ/ザ・ファースト・アベンジ...	movie-enter
...	...	...	...	...	...
7362	http://news.livedoor.com/article/detail/6530260/	2012-05-05T09:55:00+0900	好きな戦士を作ってドラゴンボールの世界で天下一武道会優勝だ！「挑戦！天下一武道会」【Andr...	どんな戦士を作るかはユーザー次第！¥n国民的人気を誇る鳥山明氏のマンガ／アニメである「DRA...	smax
7363	http://news.livedoor.com/article/detail/6681611/	2012-06-21T20:55:00+0900	NTTドコモ、GALAXY SIII SC-06DとF-09D ANTEPRIMAの発売日を...	GALAXY SIIIが6月28日、F-09D ANTEPRIMAが6月27日に発売！¥nN...	smax
7364	http://news.livedoor.com/article/detail/6856578/	2012-08-15T11:55:00+0900	NTTドコモ、Android向け「docomo Wi-Fiかんたん接続アプリ」をバージョンア...	shimajiro@mobiler¥nNTTドコモは、同社の公衆無線LANサービス「doco...	smax
7365	http://news.livedoor.com/article/detail/6678539/	2012-06-21T06:55:00+0900	NTTドコモ、PRADA Phone by LG L-02Dのデコメ絵文字popが正常に表示...	PRADA Phone by LG L-02Dにソフトウェア更新！¥nNTTドコモは20日、...	smax
7366	http://news.livedoor.com/article/detail/6869011/	2012-08-20T08:55:00+0900	NTTドコモ、公式オンラインショップでも端末複数台購入で最大10,500円／台の割り引きが受...	NTTドコモは17日、公式オンラインショップ「ドコモオンラインショップ」において端末を複数台...	smax

LDA(潜在的ディリクレ配分法)

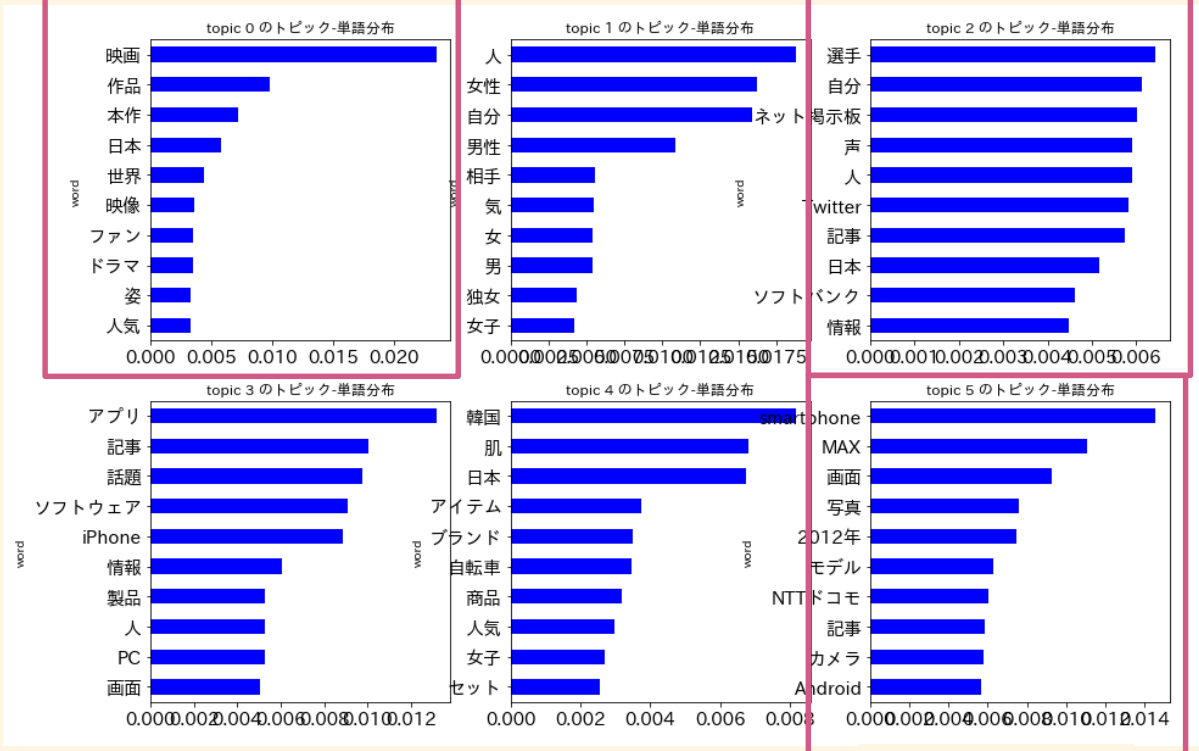
LDAの結果1

LDAでは「トピック-単語分布」と「文章-トピック分布」が出力されます。(トピック数=6)
確率分布で出力されるので、トピックごとの解釈がしやすいです。

トピック-単語分布

映画の話題

スポーツの話題



スマホの話題

文章-トピック分布

文章	トピックの分布(%)					
	0	1	2	3	4	5
女性がメイクをする上で最も気合が入るのが、アイメイク。…	22.9	**75.6**	0.4	0.4	0.4	0.4
らくらくスマートフォン F-12D特集！…	0.0	0.0	0.0	0.0	0.0	**99.9**
キムチチゲやチヂミ、サムゲタンなど、韓国料理は昔よりもはるかに身近なものになりました。…	21.0	0.2	0.2	0.2	**78.4**	0.2

LDA(潜在的ディリクレ配分法)

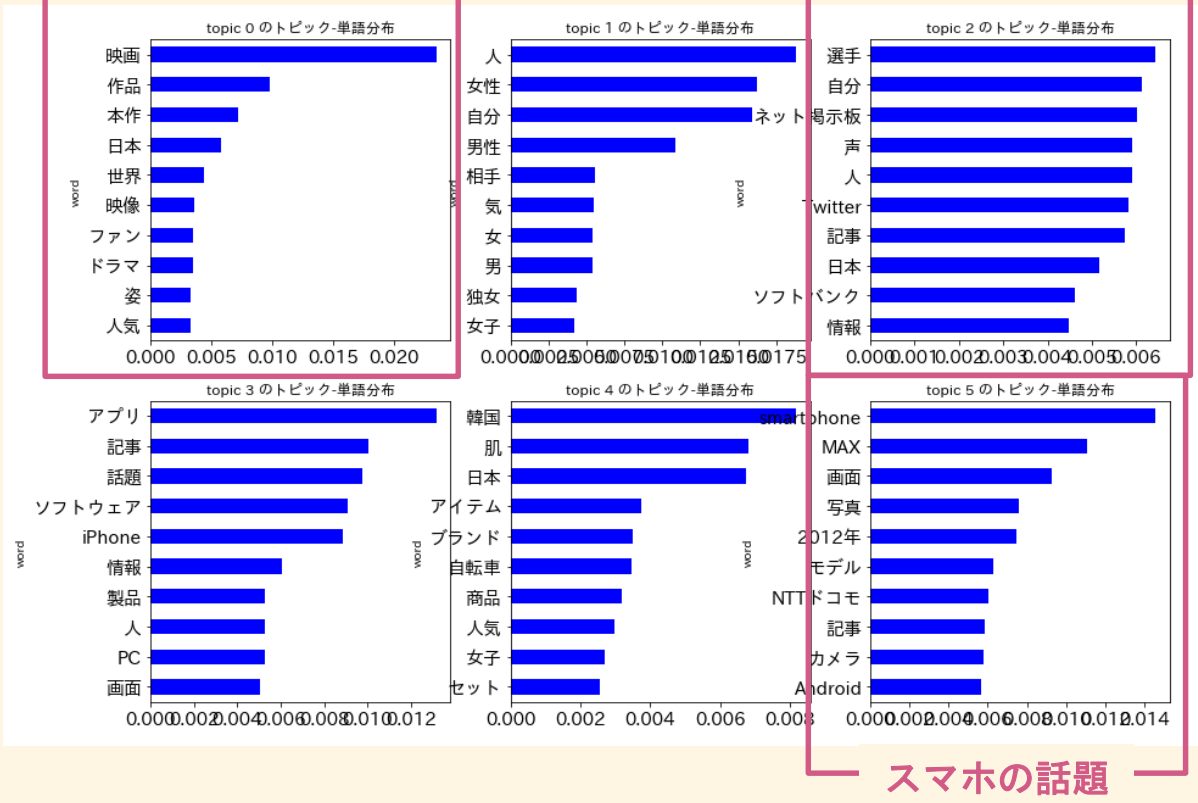
LDAの結果2

livedoorニュースコーパスにはラベルがついているので、トピックごとのラベルの分布を見てみます。トピックごとに、ラベルの特徴が見て取れます

トピック-単語分布

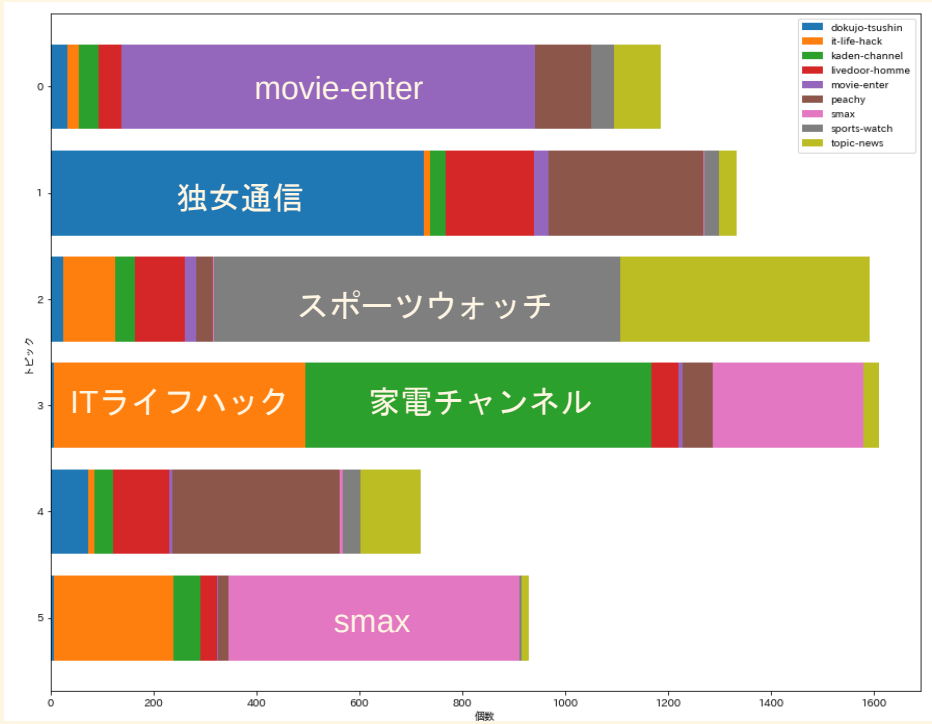
映画の話題

スポーツの話題



スマホの話題

トピックごとのラベル別度数棒グラフ



4. トピックモデルの評価

トピックモデルの評価

Nature Insight

Perplexity1

トピックモデル、特にLDAでは、分析者が最適なクラスタリング数を決める必要がある。
このとき、トピックモデルの評価としてPerplexityがあるので、それを紹介します。

以下の文章で、下線部に入る単語を考えます。

「＿＿＿＿の世界大会で、日本の選手が優勝した。」

LDAによって、この文章のトピックがスポーツであることがわかると、
下線部に入る単語はスポーツ用語に絞ることができます。

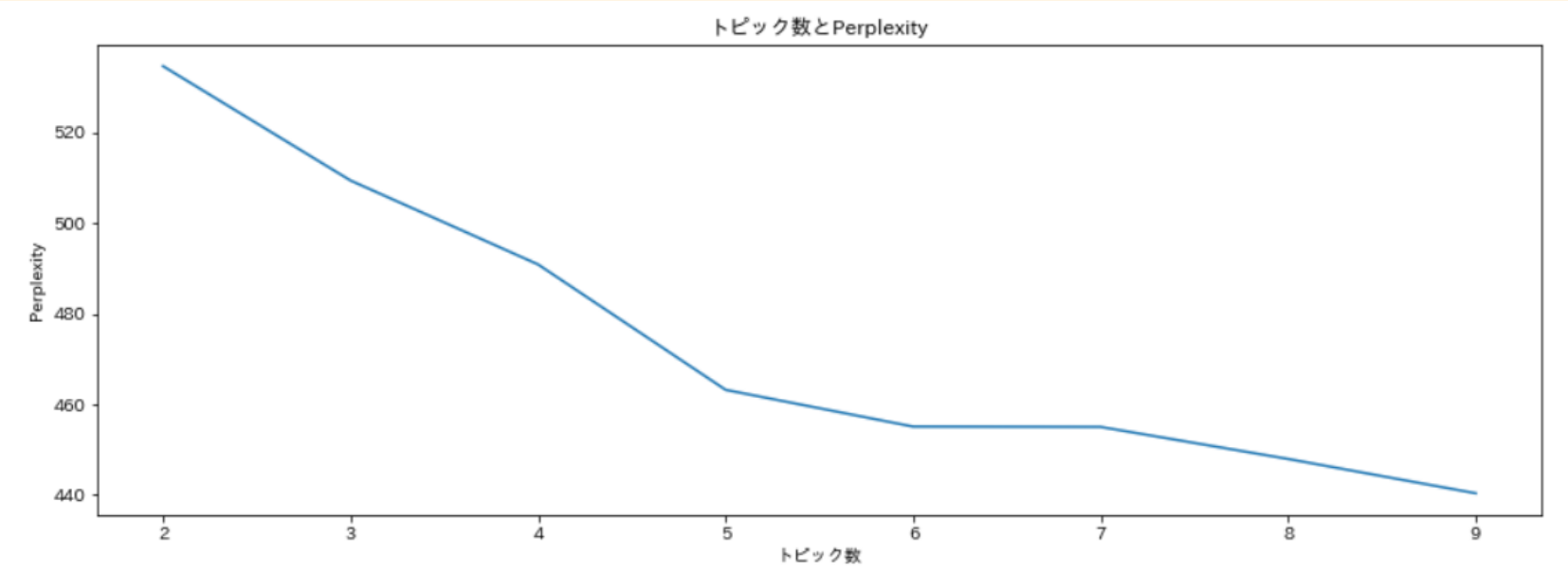
このように、LDAで適切にトピックが分かれていると、単語の選択肢が少なくなります。
この選択肢が平均的に何個かを表す指標が「Perplexity」です。

トピックモデルの評価

Perplexity2

実際にlivedoorニュースコーパスで、トピック数を変えてPerplexityを計算した結果が以下のようになります。Perplexityが緩やかになる周辺のトピック数を選ぶのが良いとされています。

(トピック数は6あたりで緩やかになっているのでこの辺りを選ぶのがよいです。)



しかし、実務では緩やかにならない場合も多いので、実際にトピック-単語分布を見てうまく分かれているかを見るほうが多いです。

5. トピックモデルの応用例

トピックモデルの応用例

Nature Insight

トピックモデルの応用例

トピックモデルには、主にクラスタリングと次元削減の2つの用途があります。

- クラスタリング

トピックモデル、特にLDAはほとんどクラスタリングを目的として使用されます。

ただし、文章以外でも、共起が重要なカウントベースの特徴量に対してクラスタリングに使用できる場合もあります。

(例：薬の服用に関する分類、カードゲームにおける使用カードの分類など)

- 次元削減

次元削減した特徴量を用いると、さまざまな機械学習タスクに応用できます。

次元削減では、トピックの解釈よりも実行速度や値の違いなどが重要になることが多いため、特異値分解ベースのLSAが使用されるケースが多いです。

(ただし、トピックの解釈が必要になる場合は、LDAが使用されることもあります。)

トピックモデルの応用例

LSAを使用した教師あり学習

LSAで次元削減した後の文章-トピック特徴量は、機械学習のINPUTとして使用できます。
入力を単語の次元数→トピックの次元数にすることで、以下のメリットがあります。

- 次元数が減り、モデルの計算効率が上がる
- トピックとして扱うため、類義性(同じ意味だが異なる単語)や多義性(同じ単語だが異なる意味)の考慮ができる

LSAを使用した機械学習の流れ

Bag-of-Words

	寿司	美味い	車	運転…
doc1	3	2	1	3
doc2	5	5	0	0
doc3	6	2	0	2
doc4	1	0	5	1

文章数×単語数

LSAで
次元削減

文章-トピック特徴量

	topic1	topic2
doc1	0.3	0.1
doc2	0.1	0.4
doc3	0.5	0.4
doc4	0.5	-1.2

文章数×トピック数

機械学習モデル

予測値

	pred
doc1	0.92
doc2	0.85
doc3	0.45
doc4	0.11

トピックモデルの応用例

Nature Insight

文章の類似度の算出

次元削減した後の特徴量からcos類似度を使用して、類似度を測ることもできます。

文章-トピック特徴量

	topic1	topic2	topic3	doc1とのcos類似度
doc1：寿司美味しい	0.9	0.1	0	1.000
doc2：寿司美味い	0.8	0.2	0	0.991
doc3：車を買う	0.1	0	0.9	0.110

同じようなトピックの分布になっている
doc1、doc2の類似度は高い

異なるトピックの分布になっている
doc1、doc3の類似度は低い

7. まとめ

トピックモデルまとめ

- 文章にはトピックがあるとして、このトピックをもとに分類する手法を「トピックモデル」という。
- トピックモデルには、LSA、PLSA、LDAがある。
- LSAは特異値分解を利用した次元削減手法であり、単語の次元をトピックの次元として表すことができる。さらに、そのトピックごとの単語の重みを出力することができる。
- LDAはLSAを確率モデルとして表した手法で、「文章-トピック分布」と「トピック-単語分布」を出力する。主にクラスタリング手法として使われることが多く、トピックの解釈も容易になっている
- 評価指標として、Perplexityがある。これは、単語の選択肢の数を表す指標であり、クラスタ数の決定の参考となる。ただし、実際にクラスタ数を決定する際は「トピック-単語分布」を見て決める場合も多い
- 次元削減の応用として、文章の類似度の算出、教師あり学習のInputとして使用などがある。

演習問題

演習問題

1. 「トピックモデル.ipynb」のLDAで、使用する品詞やパラメータを変更して、
   どのようにクラスタリング結果が変化するかみてみましょう。
2. 「トピックモデル.ipynb」を参考に、yahoo_review_trainとyahoo_review_testを
   結合したデータを使用して、LSA/LDAモデルを作成してみましょう。
   データはレビューごとですが、分類時はお店を1レコードとして扱い、お店ごとのクラスタリングをしてみましょう。(解答は演習問題2-2にあります。)
3. クラスタリング結果を解釈してみましょう。
   解釈が難しい場合には、stop_wordや品詞やパラメータをいろいろ変えて実施してみましょう。

参考文献

1. トピックモデルによる統計的潜在意味解析 ,佐藤 一誠 著 ,奥村 学 監修

東京都千代田区神田小川町３-３
HF神田小川町ビルディング５階

03-3518-6061(代)

http://www.n-insight.co.jp/